Towards a Better Understanding of Discourse: Integrating Multiple Discourse Annotation Perspectives Using UIMA

Claudiu Mihăilă*, Georgios Kontonatsios*, Riza Theresa Batista-Navarro*, Paul Thompson*, Ioannis Korkontzelos and Sophia Ananiadou
The National Centre for Text Mining,
School of Computer Science, The University of Manchester
{mihailac,kontonag,batistar,thompsop,
korkonti,ananiads}@cs.man.ac.uk

Abstract

There exist various different discourse annotation schemes that vary both in the perspectives of discourse structure considered and the granularity of textual units that are annotated. Comparison and integration of multiple schemes have the potential to provide enhanced information. However, the differing formats of corpora and tools that contain or produce such schemes can be a barrier to their integration. U-Compare is a graphical, UIMA-based workflow construction platform for combining interoperable natural language processing (NLP) resources, without the need for programming skills. In this paper, we present an extension of U-Compare that allows the easy comparison, integration and visualisation of resources that contain or output annotations based on multiple discourse annotation schemes. The extension works by allowing the construction of parallel sub-workflows for each scheme within a single U-Compare workflow. The different types of discourse annotations produced by each sub-workflow can be either merged or visualised side-by-side for comparison. We demonstrate this new functionality by using it to compare annotations belonging to two different approaches to discourse analysis, namely discourse relations and functional discourse annotations. Integrating these different annotation types within an interoperable environment allows us to study the correlations between different types of discourse and report on the new insights that this allows us to discover.

*The authors have contributed equally to the development of this work and production of the manuscript.

1 Introduction

Over the past few years, there has been an increasing sophistication in the types of available natural language processing (NLP) tools, with named entity recognisers being complemented by relation and event extraction systems. Such relations and events are not intended to be understood in isolation, but rather they are arranged to form a coherent discourse. In order to carry out complex tasks such as automatic summarisation to a high degree of accuracy, it is important for systems to be able to analyse the discourse structure of texts automatically. To facilitate the development of such systems, various textual corpora containing discourse annotations have been made available to the NLP community. However, there is a large amount of variability in the types of annotations contained within these corpora, since different perspectives on discourse have led to the development of a number of different annotation schemes.

Corpora containing discourse-level annotations usually treat the text as a sequence of coherent textual zones (e.g., clauses and sentences). One line of research has been to identify which zones are logically connected to each other, and to characterise these links through the assignment of *discourse relations*. There are variations in the complexity of the schemes used to annotate these discourse relations. For example, Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) defines 23 types of discourse relations that are used to structure the text into complex discourse trees. Whilst this scheme was used to enrich the Penn TreeBank (Carlson et al., 2001), the Penn Discourse TreeBank (PDTB) (Prasad et al., 2008) used another scheme to identify discourse relations that hold between pairs of text spans. It categorises the relations into types such as "causal", "temporal" and "conditional", which can be either explicit or implicit, depending on whether or

Proceedings of the 7th Linguistic Annotation Workshop & Interoperability with Discourse, pages 79–88,
Sofia, Bulgaria, August 8-9, 2013.

not they are represented in text using overt *discourse connectives*. In the biomedical domain, the Biomedical Discourse Relation Bank (BioDRB) (Prasad et al., 2011) annotates a similar set of relation types, whilst BioCause focusses exclusively on causality (Mihăilă et al., 2013).

A second line of research does not aim to link textual zones, but rather to classify them according to their specific function in the discourse. Examples of functional discourse annotations include whether a particular zone asserts new information into the discourse or represents a speculation or hypothesis. In scientific texts, knowing the type of information that a zone represents (e.g., background knowledge, hypothesis, experimental observation, conclusion, etc.) allows for automatic isolation of new knowledge claims (Sándor and de Waard, 2012). Several annotation schemes have been developed to classify textual zones according to their rhetorical status or general information content (Teufel et al., 1999; Mizuta et al., 2006; Wilbur et al., 2006; de Waard and Pander Maat, 2009; Liakata et al., 2012a). Related to these studies are efforts to capture information relating to discourse function at the level of events, i.e., structured representations of pieces of knowledge which, when identified, facilitate sophisticated semantic searching (Ananiadou et al., 2010). Since there can be multiple events in a sentence or clause, the identification of discourse information at the event level can allow for a more detailed analysis of discourse elements than is possible when considering larger units of text. Certain event corpora such as ACE 2005 (Walker, 2006) and GENIA-MK (Thompson et al., 2011) have been annotated with various types of functional discourse information.

It has previously been shown that considering several functional discourse annotation schemes in parallel can be beneficial (Liakata et al., 2012b), since each scheme offers a different perspective. For a common set of documents, the cited study analysed and compared functional discourse annotations at different levels of textual granularity (i.e., sentences, clauses and events), showing how the different schemes could complement each other in order to lay the foundations for a possible future harmonisation of the schemes. The results of this analysis provide evidence that it would be useful to carry out further such analyses involving other such schemes, including an investigation of how discourse relations and functional discourse annotations could complement each other, e.g., which types of functional annotations occur within the arguments of discourse relations. There are, however, certain barriers to carrying out such an analysis. For example, a comparison of annotation schemes would ideally allow the different types of annotations to be visualised simultaneously or seamlessly merged together. However, the fact that annotations in different corpora are encoded using different formats (e.g., stand-off or in-line) and different encoding schemes means that this can be problematic.

A solution to the challenges introduced above is offered by the Unstructured Information Management Architecture (UIMA) (Ferrucci and Lally, 2004), which defines a common workflow metadata format facilitating the straightforward combination of NLP resources into a workflow. Based on the interoperability of the UIMA framework, numerous researchers distribute their own tools as UIMA-compliant components (Kano et al., 2011; Baumgartner et al., 2008; Hahn et al., 2008; Savova et al., 2010; Gurevych et al., 2007; Rak et al., 2012b). However, UIMA is only intended to provide an abstract framework for the interoperability of language resources, leaving the actual implementation to third-party developers. Hence, UIMA does not explicitly address interoperability issues of tools and corpora.

U-Compare (Kano et al., 2011) is a UIMA-based workflow construction platform that provides a graphical user interface (GUI) via which users can rapidly create NLP pipelines using a drag-and-drop mechanism. Conforming to UIMA standards, U-Compare components and pipelines are compatible with any UIMA application via a common and sharable type system (i.e., a hierarchy of annotation types). In defining this type system, U-Compare promotes interoperability of tools and corpora, by exhaustively modelling a wide range of NLP data types (e.g., sentences, tokens, part-of-speech tags, named entities). This type system was recently extended to include discourse annotations to model three discourse phenomena, namely causality, coreference and meta-knowledge (Batista-Navarro et al., 2013).

In this paper, we describe our extensions to U-Compare, supporting the integration and visualisation of resources annotated according to multiple discourse annotation schemes. Our method

decomposes pipelines into parallel sub-workflows, each linked to a different annotation scheme. The resulting annotations produced by each sub-workflow can be either merged within a single document or visualised in parallel views.

2 Related work

Previous studies have shown the advantages of comparing and integrating different annotation schemes on a corpus of documents (Guo et al., 2010; Liakata et al., 2010; Liakata et al., 2012b). Guo et al. (2010) compared three different discourse annotation schemes applied to a corpus of biomedical abstracts on cancer risk assessment and concluded that two of the schemes provide more fine-grained information than the other scheme. They also revealed a subsumption relation between two schemes. Such outcomes from comparing schemes are meaningful for users who wish to select the most appropriate scheme for annotating their data. Liakata et al. (2012) underline that different discourse annotation schemes capture different dimensions of discourse. Hence, there might be complementary information across different schemes. Based on this hypothesis, they provide a comparison of three annotation schemes, namely CoreSC (Liakata et al., 2012a), GENIA-MK (Thompson et al., 2011) and DiscSeg (de Waard, 2007), on a corpus of three full-text papers. Their results showed that the categories in the three schemes can complement each other. For example, the values of the *Certainty Level* dimension of the GENIA-MK scheme can be used to assign confidence values to the Conclusion, Result, Implication and Hypothesis categories of CoreSC and DiscSeg. In contrast to previous studies, our proposed approach automatically integrates multiple annotation schemes. The proposed mechanism allows users to easily compare, integrate and visualise multiple discourse annotation schemes in an interoperable NLP infrastructure, i.e., U-Compare.

There are currently a number of freely-available NLP workflow infrastructures (Ferrucci and Lally, 2004; Cunningham et al., 2002; Schäfer, 2006; Kano et al., 2011; Grishman, 1996; Baumgartner et al., 2008; Hahn et al., 2008; Savova et al., 2010; Gurevych et al., 2007; Rak et al., 2012b). Most of the available infrastructures support the development of standard NLP applications, e.g., part-of-speech tagging, deep parsing, chunking, named entity recognition and several of them allow the representation and analysis of discourse phenomena (Kano et al., 2011; Cunningham et al., 2002; Savova et al., 2010; Gurevych et al., 2007). However, none of them has demonstrated the integration of resources annotated according to multiple annotation schemes within a single NLP pipeline.

GATE (Cunningham et al., 2002) is an open source NLP infrastructure that has been used for the development of various language processing tasks. It is packaged with an exhaustive number of NLP components, including discourse analysis modules, e.g., coreference resolution. Furthermore, GATE offers a GUI environment and wrappers for UIMA-compliant components. However, GATE implements a limited workflow management mechanism that does not support the execution of parallel or nested workflows. In addition to this, GATE does not promote interoperability of language resources since it does not define any hierarchy of NLP data types and components do not formally declare their input/output capabilities.

In contrast to GATE, UIMA implements a more sophisticated workflow management mechanism that supports the construction of both parallel and nested pipelines. In this paper, we exploit this mechanism to integrate multiple annotation schemes in NLP workflows. cTAKES (Savova et al., 2010) and DKPro (Gurevych et al., 2007) are two repositories containing UIMA-compliant components that are tuned for the medical and general domain, respectively. However, both of these repositories support the representation of only one discourse phenomenon, i.e., coreference. Argo (Rak et al., 2012a; Rak et al., 2012b) is a web-based platform that allows multiple branching and merging of UIMA pipelines. It incorporates several U-Compare components and consequently, supports the U-Compare type system.

3 A UIMA architecture for processing multiple annotation schemes

In UIMA, a document, together with its associated annotations, is represented as a standardised data structure, namely the Common Analysis Structure (CAS). Each CAS can contain any number of nested sub-CASes, i.e., *Subjects of Analysis* (*Sofas*), each of which can associate a different type of annotation with the input document. In this paper, we employ this UIMA mechanism to allow the integration and comparison of multiple

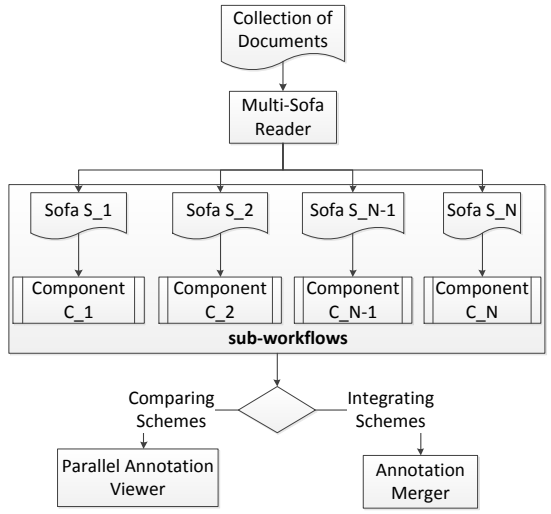

Figure 1: Integrating annotations from multiple annotation schemes in UIMA workflows

annotation schemes in a single U-Compare workflow. Assume that we have a corpus of documents which has been annotated according to n different schemes, $S_1, S_2, ..., S_{n-1}, S_n$. Also, assume that we will use a library of m text analysis components, $C_1, C_2, ..., C_{m-1}, C_m$, to enrich the corpus with further annotations.

Our implemented architecture is illustrated in Figure 1. Using multiple Sofas, we are able to split a UIMA workflow into parallel sub-workflows. Starting from a *Multi-Sofa reader*, we create n sub-workflows, i.e., Sofas, each of which is linked to a particular scheme for a different annotation type. Each sub-workflow can then apply the analysis components that are most suitable for processing the annotations from the corresponding scheme.

U-Compare offers two different modes for visualising corpora that have been annotated according to multiple schemes. In the comparison mode, the default annotation viewer is automatically split to allow annotations from different schemes to be displayed side-by-side. The second type of visualisation merges the annotations produced by the parallel sub-workflows into a single view. The most appropriate view may depend on the preferences of the user and the task at hand, e.g., identifying similarities, differences or complementary information between different schemes.

4 Application Workflows

In this section, we demonstrate two workflow applications that integrate multiple discourse annotation schemes. The first workflow exploits U-Compare's comparison mode to visualise in parallel functional discourse annotations from two schemes, namely, CoreSC (Liakata et al., 2012a) and GENIA-MK (Thompson et al., 2011). The second application integrates functional discourse annotations in the ACE 2005 corpus with discourse relations obtained by an automated tool.

4.1 Visualising functional discourse annotations from different schemes

The purpose of this workflow application is to reveal the different interpretations given by two discourse annotation schemes applied to a biomedical corpus of three full-text papers (Liakata et al., 2012b). The pipeline contains two readers that take as input the annotations (in the BioNLP Shared Task stand-off format) from the two schemes and map them to U-Compare's type system. In this way, the annotations become interoperable with existing components in U-Compare's library. U-Compare detects that the workflow contains two annotation schemes and automatically creates two parallel sub-workflows as explained earlier. Furthermore, we configure the workflow to use the comparison mode. Therefore, the annotation viewer will display the two different types of annotations based on the input schemes side-by-side. Figure 2 illustrates the parallel viewing of a document annotated according to both the CoreSC (left-hand side) and GENIA-MK (right-hand side) annotation schemes. The CoreSC scheme assigns a single category per sentence. The main clause in the highlighted sentence on the left-hand side constitutes the hypothesis that *transcription factors bind to exon-1*. Accordingly, as can be confirmed from the annotation table on the far right-hand side of the figure, the *(Hyp)othesis* category has been assigned to the sentence.

In the GENIA-MK corpus, the different pieces of information contained within the sentence have been separately annotated as structured events. One of these events corresponds to the hypothesis, but this is not the only information expressed: information about a previous experimental outcome from the authors, i.e., that exon1 is implicated in CCR3 transcription, is annotated as a sep-

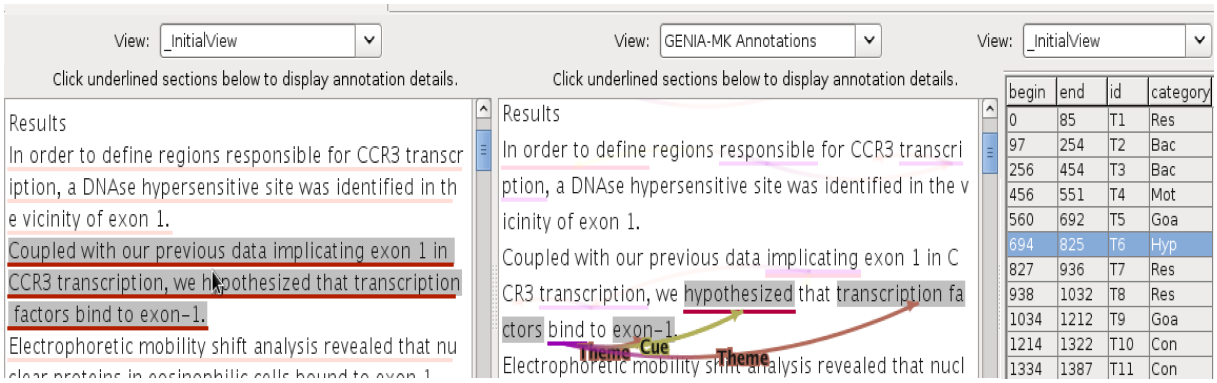

Figure 2: Comparing discourse annotations schemes in U-Compare. The pipeline uses two Sofas corresponding to the CoreSC (left panel) and GENIA-MK (right panel) schemes.

arate event. Since functional discourse information is annotated directly at the event level in the GENIA-MK corpus, the *bind* event is considered independently from the other event as representing an *Analysis*. Furthermore, the word *hypothesized* is annotated as a cue for this categorisation. There are several ways in which the annotations of the two schemes can be seen to be complementary to each other. For example, the finer-grained categorisation of analytical information in the CoreSC scheme could help to determine that the analytical *bind* event in the GENIA-MK corpus specifically represents a hypothesis, rather than, e.g., a conclusion. Conversely, the event-based annotation in the GENIA-MK corpus can help to determine exactly which part of the sentence represents the hypothesis. Furthermore, the cue phrases annotated in the GENIA-MK corpus could be used as additional features in a system trained to assign CoreSC categories. Although in this paper we illustrate only the visualisation of different types of functional discourse annotations, it is worth noting that U-Compare provides support for further processing. Firstly, unlike annotation platforms such as brat (Stenetorp et al., 2012), U-Compare allows for analysis components to be integrated into workflows in a straightforward and user-interactive manner. If, for example, it is of interest to determine the tokens (and the corresponding parts-of-speech) which frequently act as cues in *Analysis* events, syntactic analysis components (e.g., tokenisers and POS taggers) can be incorporated via a drag-and-drop mechanism. Also, U-Compare allows the annotations to be saved in a computable format using the provided Xmi Writer CAS Consumer component. This facilitates further automatic comparison of annotations.

4.2 Integrating discourse relations with functional discourse annotations

To demonstrate the integration of annotations originating from two completely different perspectives on discourse, we have created a workflow that merges traditional discourse relations with functional discourse annotations in a general domain corpus. For this application, we used the ACE 2005 corpus, which consists of 599 documents coming from broadcast conversation, broadcast news, conversational telephone speech, newswire, weblog and usenet newsgroups. This corpus contains event annotations which have been enriched by attributes such as *polarity* (positive or negative), *modality* (asserted or other), *genericity* (generic or specific) and *tense* (past, present, future or unspecified). We treat the values of these attributes as functional discourse annotations, since they provide further insight into the interpretation of the events. We created a component that reads the event annotations in the corpus and maps them to U-Compare's type system.

To obtain discourse relation annotations (which are not available in the ACE corpus) we employed an end-to-end discourse parser trained on the Penn Discourse TreeBank (Lin et al., 2012). It outputs three general types of annotations, namely, explicit relations, non-explicit relations and attribution spans. Explicit relations (i.e., those having overt discourse connectives) are further categorised into the following 16 PDTB level-2 types: Asynchronous, Synchrony, Cause, Pragmatic_cause, Contrast, Concession, Conjunction, Instantiation, Restatement, Alternative, List, Condition, Pragmatic_condition, Pragmatic_contrast, Pragmatic_concession and Excep-

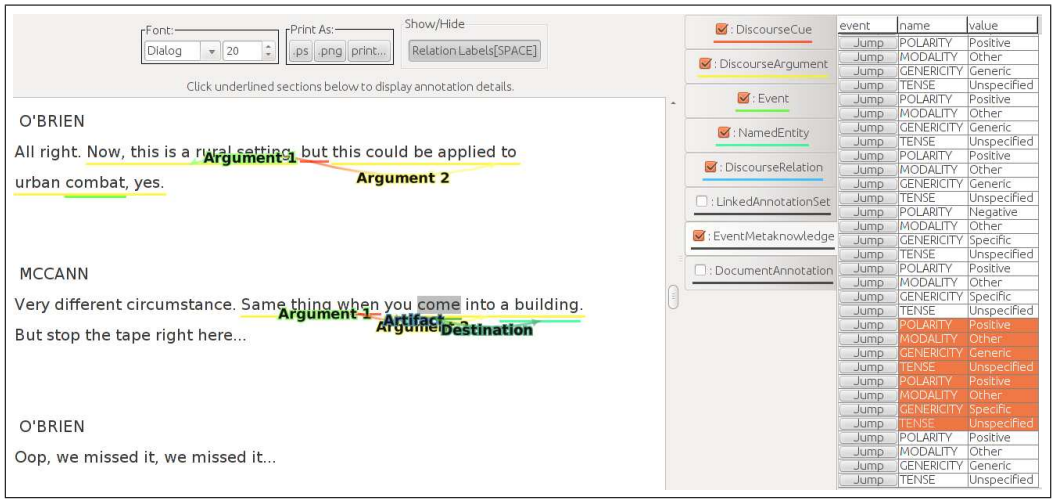

Figure 3: Integrating different discourse annotation schemes in U-Compare.

tion. Non-explicit relations, on the other hand, consist of EntRel and NoRel types, in addition to the same first 11 explicit types mentioned above.

We created a workflow consisting of the ACE corpus reader and the discourse parser (available in U-Compare as a UIMA web service). This allowed us to merge traditional discourse relations with event-based functional discourse annotations, and to visualise them in the same document (Figure 3). Furthermore, with the addition of the Xmi Writer CAS Consumer in the workflow, the merged annotations can be saved in a computable format for further processing, allowing users to perform deeper analyses on the discourse annotations. This workflow has enabled us to gain some insights into the correlations between functional discourse annotations and discourse relations.

5 Correlations between discourse relations and functional discourse annotations

Based on the merged annotation format described in the previous section, we computed cases in which at least one of the arguments of a discourse relation also contains an event. Figure 4 is a heatmap depicting the correlations between different types of discourse relations and the attribute values of ACE events that co-occur with these relations. The darker the colour, the smaller the ratio of the given discourse relation co-occurring with the specified ACE event attribute value. For instance, the *Cause* relation co-occurs mostly with *positive* events (over 95%) and the corresponding cell is a very light shade of green. These are discussed and exemplified below. In the examples, the following marking convention is used: discourse connectives are capitalised, whilst arguments are underlined. Event triggers are shown in bold, and cues relating to functional discourse categories are italicised.

For all discourse relation types, at least 50% of co-occurring events are assigned the *specific* value of the *Genericity* attribute. Specific events are those that describe a specific occurrence or situation, rather than a more *generic* situation. In general, this high proportion of *specific* events is to be expected. The types of text contained within the corpus, consisting largely of news and transcriptions of conversions, would be expected to introduce a large amount of information about specific events.

For two types of discourse relations, i.e. *Condition* and *Concession*, there are more or less equal numbers of *specific* and *generic* events. The nature of these relation types helps to explain these proportions. Conditional relations often describe how a particular, i.e., *specific*, situation will hold if some hypothetical situation is true. Since hypothetical situations do not denote specific instances, they will usually be labelled as *generic*. Concessions, meanwhile, usually describe how a specific situation holds, even though another (more generic) situation would normally hold, that would be inconsistent with this. For the *Instantiation* relation category, it may once again be expected that similar proportions of *generic* and *specific* events would co-occur within their arguments, since an instantiation describes a specific instance of a more generic situation. However, contrary to these

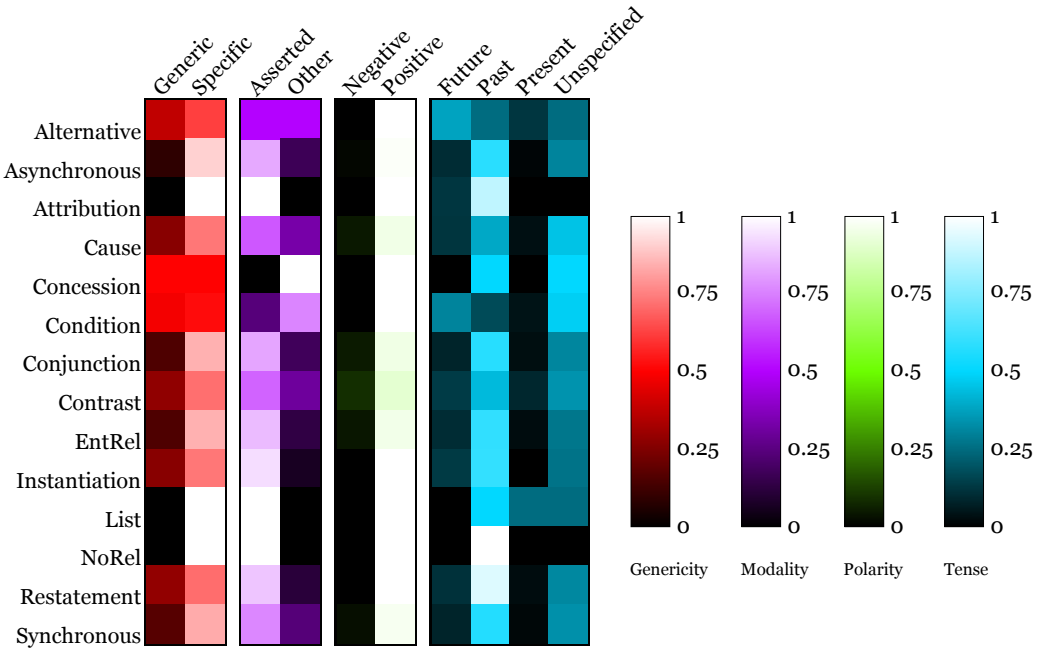

Figure 4: Heatmap showing the distribution of correlations between discourse relations and event-based functional discourse categories. A darker shade indicates a smaller percentage of instances of a discourse relation co-occurring with an event attribute.

expectations, the ratio of specific to generic events is 3:1. A reason for this is that discourse arguments corresponding to the description of a specific instance may contain several different events, as illustrated in Example (1).

(1) Toefting has been **convicted** before. In 1999 he was given a 20-day suspended **sentence** for **assaulting** a fan who berated him for playing with German club Duisburg.

In terms of the *Modality* attribute, most discourse relations correlate with definite, *asserted* events. Simillarly to the *Genericity* attribute, this can be largely explained by the nature of the texts. However, there are two relation types, i.e., *Condition* and *Consession*, which have particularly high proportions of co-occurring events whose modality is *other*. Events that are assigned this attribute value correspond to those that are *not* described as though they are real occurrences. This includes, e.g., speculated or hypothetical events. The fact that *Condition* relations are usually hypothetical in nature explain why 76% of events that co-occur with such relations are assigned the *other* value for the *Modality* attribute. Example (2) illustrates a sentence containing this relation type.

(2) And I've said many times, IF we all agreed on everything, everybody *would* want to **marry** Betty and we would really be in a mess, wouldn't we, Bob.

An even higher proportion of *Concession* relations co-occurs with events whose modality is *other*. Example (3) helps to explain this. In the first clause (the generic situation), the mention of minimising civilian casualties is only described as an *effort*, rather than a definite situation. The hedging of this generic situation is necessary in order to concede that the more specific situation described in the second clause could actually be true, i.e., that a large number of civilians have already been killed. Due to the nature of news reporting, which may come from potentially unreliable sources, the *killed* event in this second clause is also hedged, through the use of the word *reportedly*.

(3) ALTHOUGH the coalition leaders have repeatedly assured that every effort would be made to minimize civilian casualties in the current Iraq war, at least 130 Iraqi civilians have been *reportedly* **killed** since the war started five days ago.

Almost 96% of events that co-occur with arguments of discourse relations have positive polarity. Indeed, for eight relation types, 100% of the corresponding events are positive. This can partly be explained by the fact that, in texts reporting news,

there is an emphasis on reporting events that have happened, rather than events that did not happen. It can, however, be noted that events that co-occur with certain discourse relation types have a greater likelihood of having negative polarity. These relations include *Contrast* (9% of events having negative polarity) and Cause (5% negative events). Contrasts can include comparisons of positive and negative situations, as in Example (4), whilst for *Cause*s, it can sometimes be relevant to state that a particular situation caused a specific event *not* to take place, as shown in Example (5).

(4) The message from the Israeli government is that its soldiers are *not* **targeting** journalists, BUT that journalists who travel to places where there could be live fire exchange between Israeli forces and Palestinian gunmen have a responsibility to take greater precautions.

(5) His father *didn't* want to **invade** Iraq, BECAUSE of all these problems they're having now.

For most relation types, around 60% of their co-occurring events are annotated as describing *past* tense situations. This nature of newswire and conversations mean that this is largely to be expected, since they normally report mainly on events that have already happened. The proportion of events assigned the *future* tense value is highest when they co-occur with discourse relations of type *Alternative*. In this relation type, it is often the case that one of the arguments describes a possible future alternative to a current situation, as the case in Example (6). This possible information pattern for *Alternative* relations, where one of the arguments represents a currently occurring situation, would also help to explain why, even though very few events in general are annotated as *present* tense, almost 10% of events that co-occur with *Alternative* relations describe events that are currently ongoing. As for events whose *Tense* value is *unspecified*, two of the most common discourse relation types with which they occur are *Condition* and *Concession*. As exemplified above, *Condition* relations are often hypothetical in nature, meaning that no specific tense can be assigned. The generic argument of a *Concession* relation can also remain unmarked for tense. As in Example (3), it is not clear whether the effort to minimise civilian casualties has already been initiated, or will be initiated in the future.

(6) Saddam wouldn't be destroying missiles UNLESS he thought he *was going to* be **destroyed** if he didn't.

6 Conclusions

Given the level of variability in existing discourse-annotated corpora, it is meaningful for users to identify the relative merits of different schemes. In this paper, we have presented an extension of the U-Compare infrastructure that facilitates the comparison, integration and visualisation of documents annotated according to different annotation schemes. U-Compare constructs multiple and parallel annotation sub-workflows nested within a single workflow, with each sub-workflow corresponding to a distinct scheme. We have applied the implemented method to visualise the similarities and differences of two functional discourse annotation schemes, namely CoreSC and GENIA-MK. To demonstrate the integration of multiple schemes in U-Compare, we developed a workflow that merged event annotations from the ACE 2005 corpus (which include certain types of functional discourse information) with discourse relations obtained by an end-to-end parser. Moreover, we have analysed the merged annotations obtained by this workflow and this has allowed us to identify various correlations between the two different types of discourse annotations.

Based on the intuition that there is complementary information across different types of discourse annotations, we intend to examine how the integration of multiple discourse schemes, e.g., features obtained by merging annotations, affects the performance of machine learners for discourse analysis.

7 Acknowledgements

We are grateful to Dr. Ziheng Lin (National University of Singapore) for providing us with the discourse parser used for this work. This work was partially funded by the European Community's Seventh Framework Program (FP7/2007-2013) [grant number 318736 (OSSMETER)]; Engineering and Physical Sciences Research Council [grant numbers EP/P505631/1, EP/J50032X/1]; and MRC Text Mining and Screening (MR/J005037/1).

References

Sophia Ananiadou, Sampo Pyysalo, Junichi Tsujii, and Douglas B. Kell. 2010. Event extraction for systems biology by text mining the literature. *Trends in Biotechnology*, 28(7):381 – 390.

Riza Theresa B. Batista-Navarro, Georgios Kontonatsios, Claudiu Mihăilă, Paul Thompson, Rafal Rak, Raheel Nawaz, Ioannis Korkontzelos, and Sophia Ananiadou. 2013. Facilitating the analysis of discourse phenomena in an interoperable NLP platform. In *Computational Linguistics and Intelligent Text Processing*, volume 7816 of *Lecture Notes in Computer Science*, pages 559–571. Springer Berlin Heidelberg, March.

William A. Baumgartner, Kevin Bretonnel Cohen, and Lawrence Hunter. 2008. An open-source framework for large-scale, flexible evaluation of biomedical text mining systems. *Journal of biomedical discovery and collaboration*, 3:1+, January.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2001. Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue - Volume 16*, SIGDIAL '01, pages 1–10, Stroudsburg, PA, USA. Association for Computational Linguistics.

Hamish Cunningham, Diana Maynard, Kalina Bontcheva, and Valentin Tablan. 2002. GATE: an architecture for development of robust HLT applications. In *In Recent Advanced in Language Processing*, pages 168–175.

Anita de Waard and Henk Pander Maat. 2009. Epistemic segment types in biology research articles. In *Proceedings of the Workshop on Linguistic and Psycholinguistic Approaches to Text Structuring (LPTS 2009)*.

Anita de Waard. 2007. A pragmatic structure for research articles. In *Proceedings of the 2nd international conference on Pragmatic web*, ICPW '07, pages 83–89, New York, NY, USA. ACM.

David Ferrucci and Adam Lally. 2004. Building an example application with the unstructured information management architecture. *IBM Systems Journal*, 43(3):455–475.

Ralph Grishman. 1996. TIPSTER Text Phase II architecture design version 2.1p 19 june 1996. In *Proceedings of the TIPSTER Text Program: Phase II*, pages 249–305, Vienna, Virginia, USA, May. Association for Computational Linguistics.

Yufan Guo, Anna Korhonen, Maria Liakata, Ilona Silins Karolinska, Lin Sun, and Ulla Stenius. 2010. Identifying the information structure of scientific abstracts: An investigation of three different schemes. In *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing*, pages 99–107. Association for Computational Linguistics.

Iryna Gurevych, Max Mühlhäuser, Christof Müller, Jürgen Steimle, Markus Weimer, and Torsten Zesch. 2007. Darmstadt knowledge processing repository based on UIMA. In *Proceedings of the First Workshop on Unstructured Information Management Architecture at Biannual Conference of the GSCL*.

Udo Hahn, Ekaterina Buyko, Rico Landefeld, Matthias Mühlhausen, Michael Poprat, Katrin Tomanek, and Joachim Wermter. 2008. An overview of JCoRe, the JULIE lab UIMA component repository. In *LREC'08 Workshop 'Towards Enhanced Interoperability for Large HLT Systems: UIMA for NLP'*, pages 1–7, Marrakech, Morocco, May.

Yoshinobu Kano, Makoto Miwa, Kevin Cohen, Lawrence Hunter, Sophia Ananiadou, and Jun'ichi Tsujii. 2011. U-Compare: A modular NLP workflow construction and evaluation system. *IBM Journal of Research and Development*, 55(3):11.

Maria Liakata, Simone Teufel, Advaith Siddharthan, and Colin Batchelor. 2010. Corpora for the conceptualisation and zoning of scientific papers. In *Proceedings of LREC*, volume 10.

Maria Liakata, Shyamasree Saha, Simon Dobnik, Colin Batchelor, and Dietrich Rebholz-Schuhmann. 2012a. Automatic recognition of conceptualization zones in scientific articles and two life science applications. *Bioinformatics*, 28(7):991–1000.

Maria Liakata, Paul Thompson, Anita de Waard, Raheel Nawaz, Henk Pander Maat, and Sophia Ananiadou. 2012b. A three-way perspective on scientic discourse annotation for knowledge extraction. In *Proceedings of the ACL Workshop on Detecting Structure in Scholarly Discourse (DSSD)*, pages 37–46, July.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2012. A PDTB-styled end-to-end discourse parser. *Natural Language Engineering*, FirstView:1–34, 10.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical Structure Theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

Claudiu Mihăilă, Tomoko Ohta, Sampo Pyysalo, and Sophia Ananiadou. 2013. BioCause: Annotating and analysing causality in the biomedical domain. *BMC Bioinformatics*, 14(1):2, January.

Yoko Mizuta, Anna Korhonen, Tony Mullen, and Nigel Collier. 2006. Zone analysis in biology articles as a basis for information extraction. *International Journal of Medical Informatics*, 75(6):468 – 487. Recent Advances in Natural Language Processing for Biomedical Applications Special Issue.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odjik, Stelios

Piperidis, and Daniel Tapias, editors, *In Proceedings of the 6th International Conference on language Resources and Evaluation (LREC)*, pages 2961–2968.

Rashmi Prasad, Susan McRoy, Nadya Frid, Aravind Joshi, and Hong Yu. 2011. The biomedical discourse relation bank. *BMC Bioinformatics*, 12(1):188.

Rafal Rak, Andrew Rowley, and Sophia Ananiadou. 2012a. Collaborative development and evaluation of text-processing workflows in a UIMA-supported web-based workbench.

Rafal Rak, Andrew Rowley, William Black, and Sophia Ananiadou. 2012b. Argo: an integrative, interactive, text mining-based workbench supporting curation. *Database: The Journal of Biological Databases and Curation*, 2012.

Ágnes Sándor and Anita de Waard. 2012. Identifying claimed knowledge updates in biomedical research articles. In *Proceedings of the Workshop on Detecting Structure in Scholarly Discourse*, ACL '12, pages 10–17, Stroudsburg, PA, USA. Association for Computational Linguistics.

Guergana Savova, James Masanz, Philip Ogren, Jiaping Zheng, Sunghwan Sohn, Karin Kipper-Schuler, and Christopher Chute. 2010. Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513.

Ulrich Schäfer. 2006. Middleware for creating and combining multi-dimensional nlp markup. In *Proceedings of the 5th Workshop on NLP and XML: Multi-Dimensional Markup in Natural Language Processing*, pages 81–84. ACL.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations Session at EACL 2012*, Avignon, France, April. Association for Computational Linguistics.

Simone Teufel, Jean Carletta, and Marc Moens. 1999. An annotation scheme for discourse-level argumentation in research articles. In *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*, EACL '99, pages 110–117, Stroudsburg, PA, USA. Association for Computational Linguistics.

Paul Thompson, Raheel Nawaz, John McNaught, and Sophia Ananiadou. 2011. Enriching a biomedical event corpus with meta-knowledge annotation. *BMC Bioinformatics*, 12(1):393.

Christopher Walker. 2006. ACE 2005 Multilingual Training Corpus.

W John Wilbur, Andrey Rzhetsky, and Hagit Shatkay. 2006. New directions in biomedical text annotation: definitions, guidelines and corpus construction. *BMC Bioinformatics*, 7(1):356.